

RECEIVED
CENTRAL FAX CENTER

This listing of claims will replace all prior versions
and listings, of claims in the application:

MAY 27 2008

Claims 1-16 (canceled)

1 Claim 17 (currently amended): A computer system for
2 building a lexicon for use in capitalization
3 correction for unstructured excerpts, comprising:
4 at least one processing unit;
5 at least one storage device being coupled with
6 the at least one processing unit and storing program
7 code;
8 a ripper adapted to assemble a list of word sets
9 from unstructured content, at least one of the word
10 sets comprising a word and at least two non-standard
11 capitalization variations for the word; and
12 an aggregator adapted to aggregate at least one
13 of the word sets, the aggregator including
14 an analyzer adapted to identify non-standard
15 capitalization variations based on at least one
16 criteria; and
17 a non-standard capitalization selector
18 adapted to select at least one of the identified
19 non-standard capitalization variations within one
20 of the at least one word sets, and adding the
21 selected at least one of the identified
22 non-standard capitalization variations to the
23 lexicon, wherein the lexicon includes records,
24 each record including a word, wherein the lexicon
25 is indexed by the words included in the records,
26 and wherein at least one of the records includes

27 more than one non-standard capitalization
28 variation.

1 Claim 18 (currently amended): A computer system
2 according to Claim 17, further comprising:
3 a tokenizer adapted to tokenize the excerpt into
4 the one or more words and one or more punctuation
5 marks.

1 Claim 19 (currently amended): A computer system
2 according to Claim 18, wherein hyphenated words are
3 split into a plurality of the words.

Claim 20 (canceled)

1 Claim 21 (currently amended): A computer system
2 according to Claim 17, wherein at least one of the
3 non-standard capitalization variations occurs in an
4 excerpt having fewer than half of individual letters
5 provided in uppercase.

1 Claim 22 (currently amended): A computer system
2 according to Claim 17, further comprising:
3 a normalizer adapted to normalize a plurality of
4 the words extracted relative to a source of the
5 unstructured excerpt.

1 Claim 23 (currently amended): A computer system
2 according to Claim 17, wherein non-standard
3 capitalization variations that are identified based on

4 one or more criteria comprise only those non-standard
5 capitalization variations having at least four
6 occurrences.

1 Claim 24 (currently amended): A computer system
2 according to Claim 17, wherein at least one of the
3 non-standard capitalization variations has any
4 individual letter other than the first individual
5 letter provided in uppercase.

Claim 25 (canceled)

1 Claim 26 (currently amended): A computer system
2 according to Claim 17, further comprising:
3 a validator adapted to apply implicit rules for
4 capitalization, and skipping each of the non-standard
5 capitalization variations subject to at least one such
6 implicit rule.

1 Claim 27 (currently amended): A computer system
2 according to Claim 26, wherein the implicit rules
3 comprise skipping each of the non-standard
4 capitalization variations based on position within a
5 sentence or phrase.

1 Claim 28 (currently amended): A computer system
2 according to Claim 26, wherein the implicit rules
3 comprise at least one of (A) the non-standard
4 capitalization variation being a number, (B) the
5 non-standard capitalization variation having no
6 vowels, and (C) the non-standard capitalization

7 variation constituting at least one of an article,
8 conjunction and preposition.

1 Claim 29 (currently amended): A computer system
2 according to Claim 26, wherein the implicit rules
3 comprise normalizing a number of occurrences for each
4 of the non-standard capitalization variations relative
5 to a source of the non-standard capitalization
6 variations.

1 Claim 30 (currently amended): A computer system
2 according to Claim 26, wherein each of the word sets
3 includes a word and at least one non-standard
4 capitalization variation, each of the at least one
5 non-standard capitalization variation including a
6 frequency of occurrence count.

1 Claim 31 (currently amended): A computer system
2 according to Claim 17, further comprising:
3 a hash table maintaining the lexicon.

1 Claim 32 (currently amended): A computer system
2 according to Claim 31,
3 wherein the hash table is indexed by words.

Claims 33-50 (canceled)

1 Claim 51 (previously presented): A computer-implemented
2 method comprising:
3 a) generating a plurality of word sets from a text
4 corpus, at least one of the words sets including

- 5 - a word identified from the text corpus,
- 6 - at least one non-standard capitalization
- 7 variation of the word included in the word set,
- 8 and
- 9 - a frequency of occurrence of each of the at
- 10 least one non-standard capitalization variation
- 11 of the word included in the word set;
- 12 b) generating a lexicon using the generated
- 13 plurality of word sets, wherein the lexicon
- 14 includes, for each of a plurality of words, at least
- 15 one capitalization variation identified using at
- 16 least one criteria, wherein at least one of the
- 17 words of the lexicon includes more than one
- 18 non-standard capitalization variation identified
- 19 using the at least one criteria; and
- 20 c) storing the generated lexicon.

1 Claim 52 (previously presented): The
2 computer-implemented method of claim 51 wherein a
3 non-standard capitalization variation is identified using
4 the at least one criteria only if it occurs at least four
5 times in the text corpus.

1 Claim 53 (previously presented): The
2 computer-implemented method of claim 51 further
3 comprising:
4 d) accepting a word having a capitalization
5 defining which, if any, of the characters of the
6 word are capitalized; and
7 e) performing a capitalization correction function
8 on the word using the generated lexicon.

1 Claim 54 (previously presented): The
2 computer-implemented method of claim 53 wherein the act
3 of performing a capitalization correction function
4 includes
5 - determining if the capitalization of the
6 word matches a capitalization variation in the
7 lexicon, and
8 - not changing the capitalization of the word
9 if it was determined to match a capitalization
10 variation in the lexicon.

1 Claim 55 (previously presented): The
2 computer-implemented method of claim 53 wherein the act
3 of performing a capitalization correction function
4 includes
5 - determining if the capitalization of the
6 word matches a non-standard capitalization
7 variation in the lexicon, which non-standard
8 capitalization variation meets a frequency
9 criteria, and
10 - not changing the capitalization of the word
11 if it was determined to match a non-standard
12 capitalization variation in the lexicon.

1 Claim 56 (previously presented): Apparatus comprising:
2 a) means for generating a plurality of word sets
3 from a text corpus, at least one of the word sets
4 including
5 - a word identified from the text corpus,

- 6 - at least one non-standard capitalization
7 variation of the word included in the word set,
8 and
9 - a frequency of occurrence of each of the at
10 least one non-standard capitalization variation
11 of the word included in the word set; and
12 b) means for generating a lexicon using the
13 generated plurality of word sets, wherein the
14 lexicon includes, for each of a plurality of words,
15 at least one capitalization variation identified
16 using at least one criteria, wherein at least one of
17 the words of the lexicon includes more than one
18 non-standard capitalization variation identified
19 using the at least one criteria.

1 Claim 57 (previously presented): The apparatus of claim
2 56 wherein a non-standard capitalization variation is
3 identified using the at least one criteria only if it
4 occurs at least four times in the text corpus.

1 Claim 58 (previously presented): The apparatus of claim
2 56 further comprising:
3 c) means for accepting a word having a
4 capitalization defining which, if any, of the
5 characters of the word are capitalized; and
6 d) means for performing a capitalization correction
7 function on the word using the generated lexicon.

1 Claim 59 (previously presented): The apparatus of claim
2 58 wherein the means for performing a capitalization
3 correction function

- 4 - determine if the capitalization of the word
- 5 matches a capitalization variation in the
- 6 lexicon, and
- 7 - do not change the capitalization of the word
- 8 if it was determined to match a capitalization
- 9 variation in the lexicon.

1 Claim 60 (previously presented): The apparatus of claim
2 58 wherein the means for performing a capitalization
3 correction function

- 4 - determine if the capitalization of the word
- 5 matches a capitalization variation in the
- 6 lexicon, which capitalization variation meets a
- 7 frequency criteria, and
- 8 - do not change the capitalization of the word
- 9 if it was determined to match a capitalization
- 10 variation in the lexicon.